Grupo Godó

Big Data Engineer Test

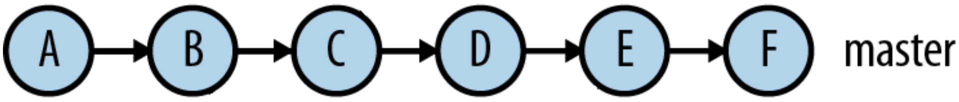
Please try to answer *all* the questions to the best of your skills. A partially correct response is better than a question not answered.
All questions can answered in English, Catalan or Spanish.

1 Cloud Computing: Amazon Web Services

- Enumerate the main differences between an *HDFS* and *AWS S3* storage systems.
- What is the *AWS IAM* service and how can be used?
- What is the *Consistent View* feature and what problem does it solve?

2 Miscellaneous

- Briefly describe the different types of NoSQL databases and its usages
- What is the difference between `Gitflow` and `Trunk Development`? Which one is better?
- What is HEAD on git? Giving the following picture what happen if the HEAD is moved to commit D? What will happen if HEAD was moved to commit D and master branch did not exist?



3 Programming

Not only the correctness of the provided solutions will be taken into account but also the quality of the responses.

For each question provide a brief reasoning explaining your strategy used to solve problem, your code (all we could need to execute it) and a sample of the output.

3.1 Java & Spark

Given the following dataset with crime entries, answer the following questions:

1. For each district and each weapon list committed crimes.
2. List most conflicting zones.
3. Provide an algorithm to detect similar crimes.

3.2 UNIX Tools & Python

Given the following file containing notable historic events, answer the following questions:

- Transform provided file to a valid JSON file by converting each event observation to a new object item of a new array.

```
{..., event: { A }, event: { B } } --> {..., event: [ { A }, { B } ] }
```

- Compute the total elements of the newly created `event` array
- Create a JSON subset containing only the array's event entries whose key `description` contains an specific given word, e.g. Barcelona, Santiago, etc.

Given the following Tuenti challenge

- Solve the problem using Python.